**Text Mining**

**Pendahuluan**

Text Mining (penambangan teks) merupakan salah satu cara yang biasa digunakan pada DataMining (penambangan data). Text Mining adalah salah satu teknik penambangan untuk jenis data yang berupa teks, dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

Pada text mining terdapat tahapan-tahapan yang dilakukan untuk memisahkan kata-kata menjadi kata dasar dari sebuah kalimat yang disebut dengan tahapan pre-processing(Mitchell1997). Tahapan-tahapan pre-processing tersebut adalah sebagai berikut:

* Case Folding

Case Folding adalah tahapan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf “a” sampai dengan “z” saja yang dapat diterima. Contoh proses case folding seperti pada Gambar 1.1. Semua huruf besar pada bagian “sebelum case folding” diubah menjadi huruf kecil. Hasilnya seperti pada bagian “hasil case folding”.

* Tokenizing

Tokenizing adalah tahapan dimana adanya pemotongan string masukan berdasarkan kata yang menyusunnya. Contoh proses tokenizing seperti pada gambar 1.2. Semua kata pada bagian “sebelum tokenizing” dipotong-potong sesuai panjang huruf dalam setiap kata yang ditandai dengan spasi. Hasilnya adalah seperti pada bagian “Hasil tokenizing”.

* Filtering

Filtering adalah tahapan mengambil kata-kata yang penting dari hasil token sebelumnya. Di dalam tahapan ini dapat digunakan algoritma pembuangan kata yangkurang penting (stoplist) atau menyimpan kata penting (wordlist). Contoh proses filering seperti pada Gambar 1.3. Semua kata yang tidak penting pada bagian “sebelum filtering” dibuang (katadan). Hasilnya adalah seperti pada bagian “Hasil filtering”. Kata yang tidak penting (stoplist) disimpan dalam satu tempat khusus untuk digunakan pada saat proses perbandingan kata dalam dokumen uji dilakukan. Apabila dalam dokumen uji terdapat kata yang sama dengan kata yang tersimpan dalam stoplist, maka kata tersebut di buang. Kata yang tidak ada dalam stoplist dijadikan kata “hasil filtering”.

* Analyzing

Analyzing merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata pada dokumen yang ada. Tabel pertama menampilkan jumlah kemunculan kata dalam dokumen. Contoh table1.1. Dalam dokumen 1 terdapat 1 kata program, 1 kata studi, 1 kata ilmu, 2 kata computer, 3 kata informatika. Tabel kedua menampilkan jumlah dokumen yang mengandung kata kunci. Contoh table 1.2., kata tumpang muncul dalam 1 dokumen, kata tindih muncul dalam 2 dokumen, kata tugas muncul dalam 1 dokumen, kata jalan muncul dalam 4 dokumen, kata penyatuan muncul dalam 1 dokumen, dan kata sistem muncul dalam 3 dokumen.